Predicting wine quality based on physicochemical attributes

Author

Zhihao Xie, Natalie Trusdell, Oswin Gan & Sam Lokanc

Published

December 6, 2025

Table of contents

Introduction

Quality assessment is an essential step in the wine making process as it helps ensure consistency and consumer satisfaction. The quality of wine has historically been evaluated by wine makers and sommeliers, who leverage their training and expertise to determine if a product meets acceptable standards Langstaff (2010). While there have been advancements in the detection of the physicochemical properties of wine and how they relate to flavour Polášková, Herszage, and Ebeler (2008), quality assessment remains a subjective metric.

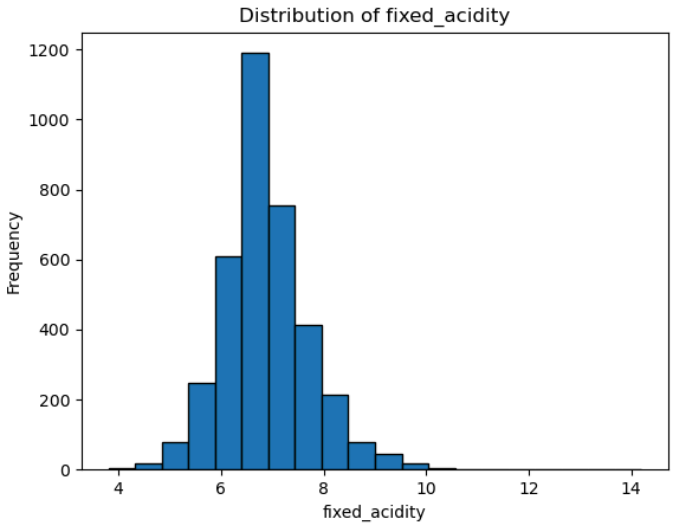
Here we ask if wine quality can be predicted using machine learning. Our focus is to create an interpretable linear regression model trained on physicochemical wine data to perform this task. The utility of such a model is two-fold as it can be used as a tool to help less experienced wine makers and sommeliers get a sense of the quality of a given wine. Furthermore, the model would help to provide a more objective framework for wine assessment through interpretation of its coefficients. To achieve this, we will train our model on data related to vinho verde white wines from the Minho region of Portugal which contains the physicochemical and sensory data of 4898 wines Cortez et al. (2009).

Methods

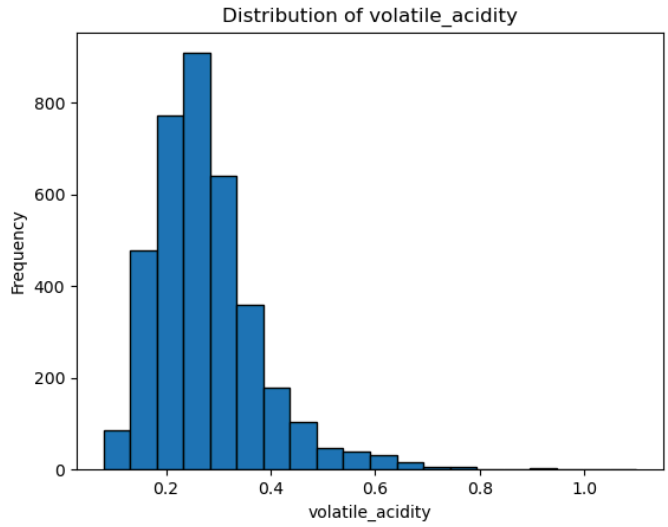
Data

The data was obtained from the UC Irvine Machine Learning Repository Dua and Graff (2017). The dataset contains physicochemical and sensory data related to 4898 white wines Cortez et al. (2009). The physicochemical data include: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, alcohol, and sulfates. These physicochemical data are emperically measured and will be used as features for our predictive model. The sensory data contained in the data set refers to the quality of the wine, which was subjectively measured by wine experts and represented as a score out of ten. This will be the target of our predictive model.

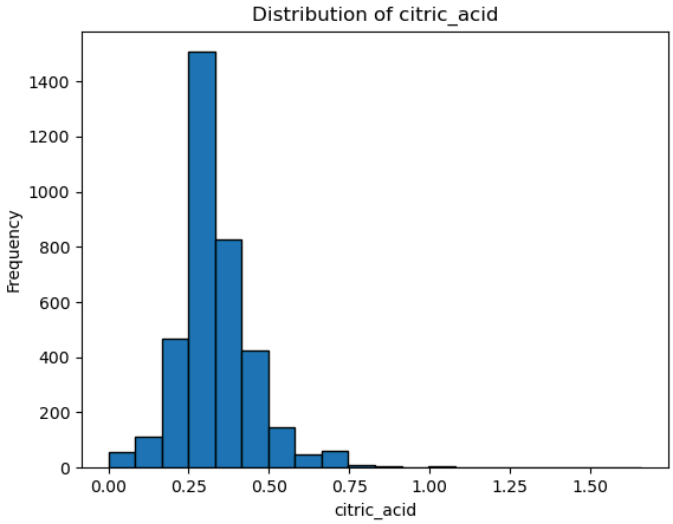
The distributions of the features can be seen in Figure 3. By taking a look at the individual distributions of the variables, we can see that the majority of them are approximately normal with some degree of right-skew. This tells us that most of the values tend to cluster around the average, possibly an industry standard, with the more extreme values tending to fall above the average rather than below it. The distributions of residual sugar and alcohol content break this pattern as the residual sugar plot follows a more exponential pattern, and the distribution of alcohol content is still approximately normal but also shows a tendency towards uniformity.



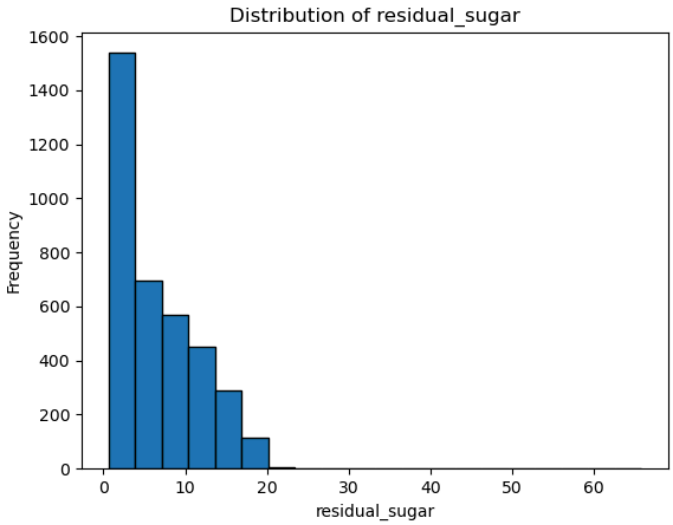
(a) Fixed Acidity



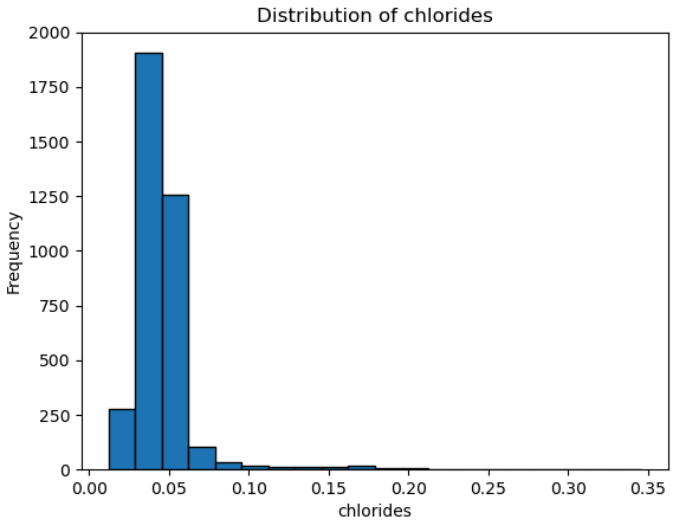
(b) Volatile Acidity



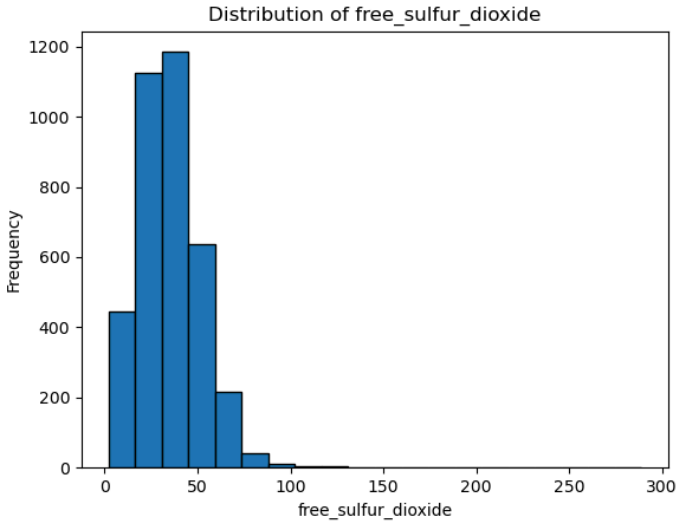
(c) Citric Acid



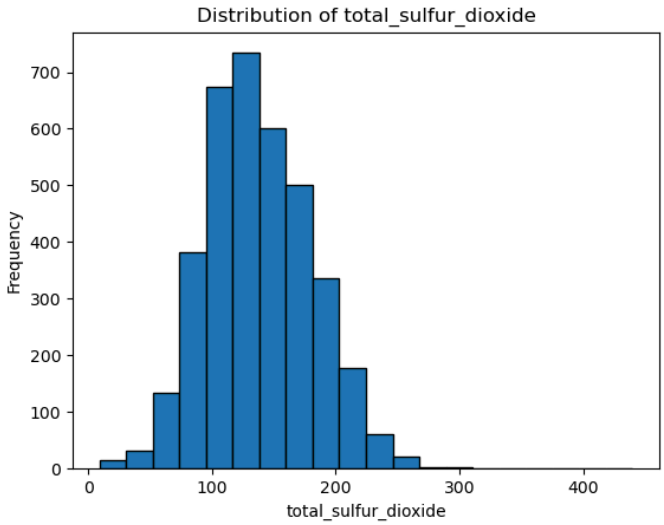
(d) Residual Sugar



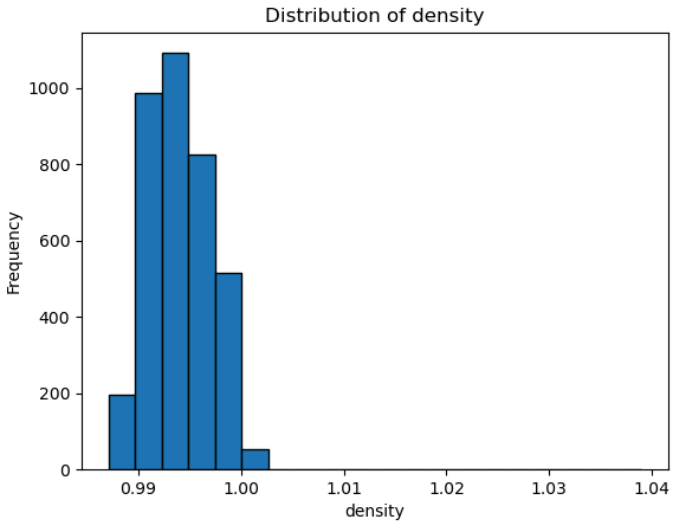
(e) Chlorides



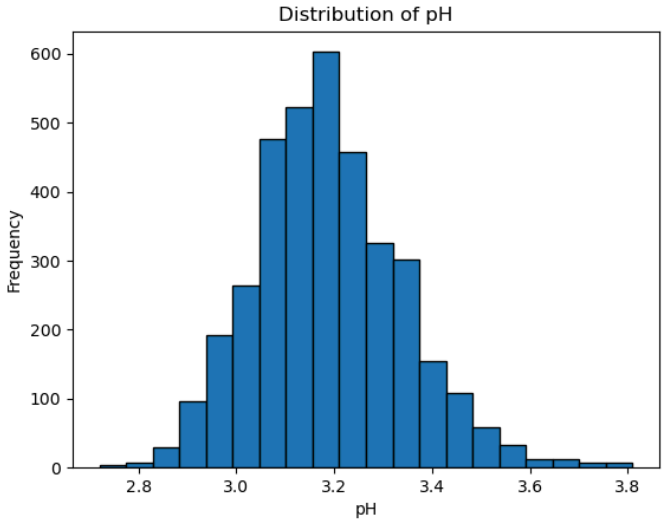
(f) Free Sulfur Dioxide



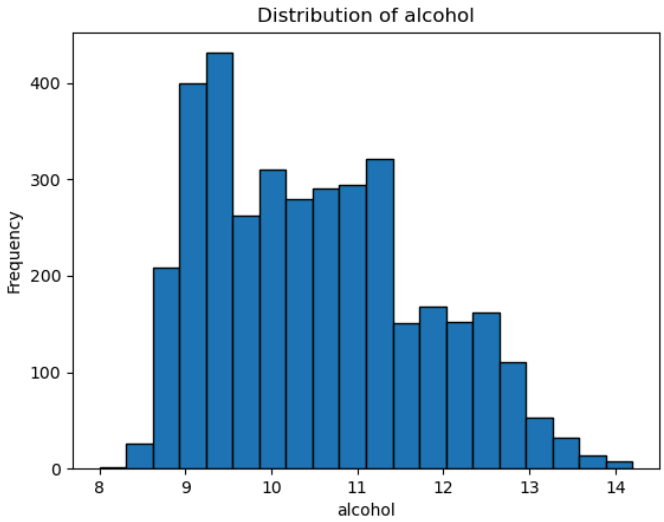
(g) Total Sulfur Dioxide



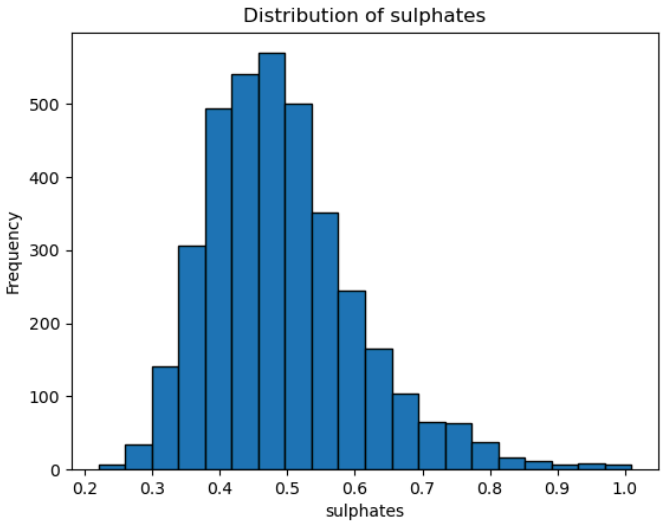
(h) Density



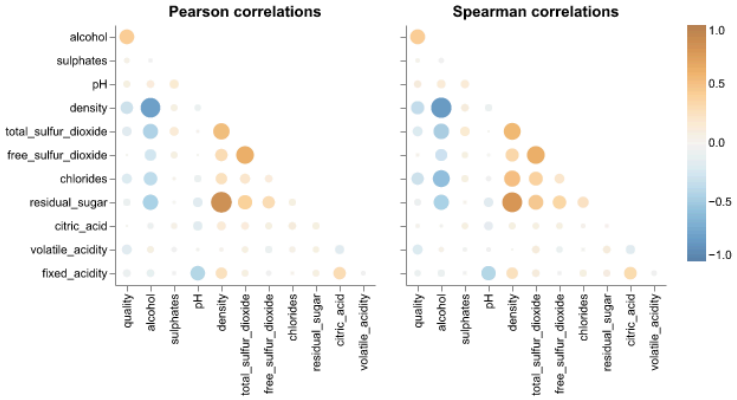
(i) pH



(j) Alcohol



(k) Sulphates



(l) Correlation

Figure 1: Physicochemical Feature Distributions and Feature Correlation

Analysis

For exploratory purposes, we first create an ordinary least squares linear regression model including all predictors:

OLS Regression Results

Dep. Variable:	quality	**R-squared:**	0.272
Model:	OLS	**Adj. R-squared:**	0.270
Method:	Least Squares	**F-statistic:**	124.6

Figure 2: OLS Regression Results

This model (Figure 2) has an R-squared value of 0.272, meaning that the model explains roughly 27.2% of variability in the data. The R-squared value is close to that of the adjusted R-squared, suggesting most predictors contribute explanatory power to the model. The p-value of the F-statistics is small, indicating that at least one predictor has a statistically significant assocation with the response. The output reveals that the condition number of the model is large, possibly indicating strong multicollinearity. Thus, we will opt for a Ridge Regression to induce an L2 penalty on correlated variables.

Table 1: Ridge Regression Analysis Results

(a) Ridge Features and Coefficients

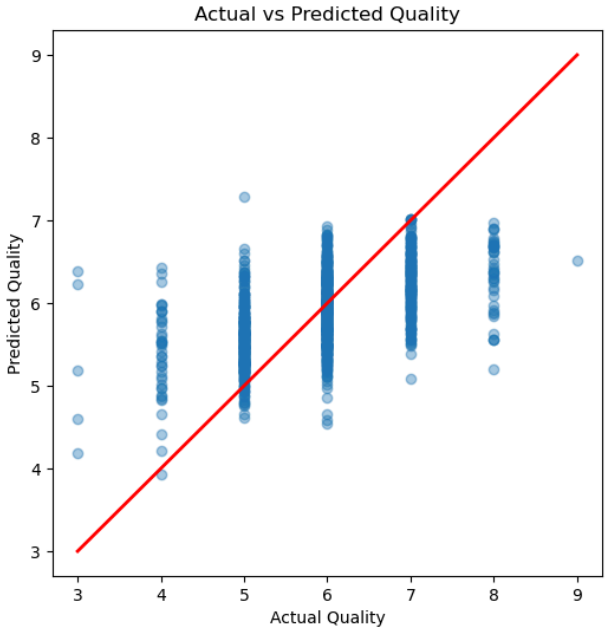
feature	coefficient
residual_sugar	0.3865
alcohol	0.2668
pH	0.0942
sulphates	0.0637
free_sulfur_dioxide	0.0547
fixed_acidity	0.0426
citric_acid	0.0048
chlorides	0
total_sulfur_dioxide	−0.0137
volatile_acidity	−0.193
density	−0.3967

(b) Ridge Regression Results

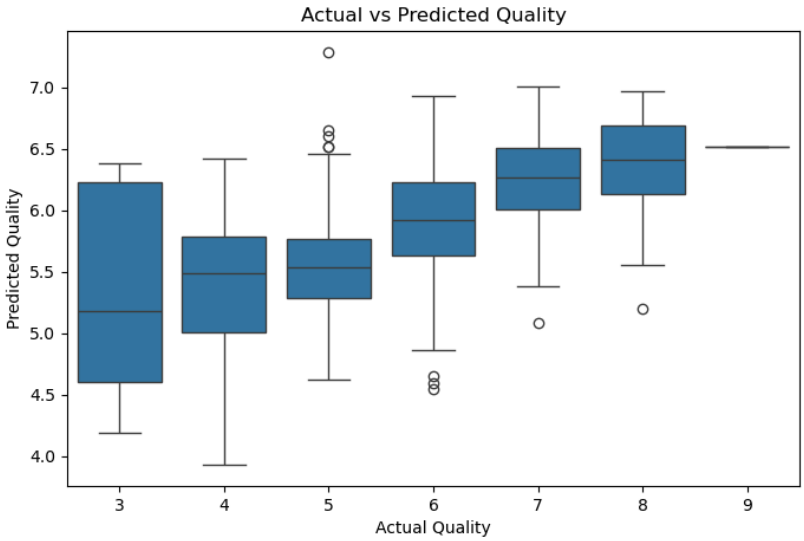
Metric	Value
RMSE	0.7357
MAE	0.5718
R^2	0.309

```
| Mean Actual    |  5.8776 |
| Mean Predicted |  5.8667 |
```

Results and Discussion



(a) Scatterplot of Actual versus Predicted Quality Scores



(b) Barplot of Actual versus Predicted Quality Scores

Figure 3

When we look at the plot for actual versus predicted quality (Figure 3 (a)), we can see that our model had a tendency to predict the middling quality ratings, suggesting the model was biased by the high frequency of quality ratings of 5-7 in our data. While our ridge regression model did tend to predict higher quality values when the actual quality value was higher, the average prediction stayed between ~5 and 6.5 for all actual quality values.

After running our ridge regression analysis, the variables that the model found most informative for predicting the quality rating were the residual sugar and alcohol content (Table 1), with a higher value of either being correlated on average with a higher predicted quality rating. Intuitively this makes sense, as wines with a higher alcohol content are considered to have more body and a richer taste, but there is also a tradeoff between alcohol content and residual sugar (Copestake (2025)). The tradeoff was hinted at in our earlier EDA, as a higher alcohol content was associated with a lower residual sugar content. The alcohol content in wines is derived from a longer fermentation process, but this fermentation also consumes more of the sugar, so it's difficult to attain high levels of both.

These findings indicate that although taste is subjective, tasters have a tendency to prefer white wines that maximize the residual sugar and alcohol content, resulting in a full-bodied wine that is not too dry. Going forward, this could prompt winemakers to experiment with grapes with a higher sugar content and fermentation techniques that try to maximize the alcohol content while minimizing the sugar consumed during the process. This also leads us to questions about how aware sommeliers are of their preferences while judging the quality of wine and how closely the quality rating matches the average wine consumer's preferences. Do sommeliers just tend to enjoy sweeter, full-bodied wines, or is that their personal opinion due to their experience? More research could be done into this topic.

References

Copestake, Nicole. 2025. "How Much Alcohol Is in Wine? A Complete Guide." Coravin. 2025. https://www.coravin.ca/blogs/community/wine-101-how-much-alcohol-is-in-wine.

Cortez, Paulo, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. 2009. “Modeling Wine Preferences by Data Mining from Physicochemical Properties.” *Decision Support Systems* 47 (4): 547–53.
Dua, Dheeru, and Casey Graff. 2017. “UCI Machine Learning Repository.” University of California, Irvine, School of Information; Computer Sciences. http://archive.ics.uci.edu/ml.
Langstaff, SA. 2010. “Sensory Quality Control in the Wine Industry.” In *Sensory Analysis for Food and Beverage Quality Control*, 236–61. Elsevier.
Polášková, Pavla, Julian Herszage, and Susan E Ebeler. 2008. “Wine Flavor: Chemistry in a Glass.” *Chemical Society Reviews* 37 (11): 2478–89.